Wrangle Report

1.Gathering Data for this Project

This project involved gathering of data from three different sources as listed below. For each of the data source a different method of data gathering was used namely: -

- Importing data via csv
- Using requests to download data off internet
- Scrape data from an API

This was challenging and fun at the same time.

Three data sources

Enhanced Twitter Archive

The ‘WeRateDogs’ Twitter archive provided by Udacity. This contains basic tweet data for all 5000+ of their tweets, but not everything. I manually downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

Image Predictions File

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: image_predictions.tsv

Data via the Twitter API

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2.Assessing data

The data we get is seldom in standard formats or to the way we want it. After gathering the required data, I came up with the following issues with it: -

Issue number	Data Frame	Issue	Type of Issue
1		Merging 3 data sets into master dataframe	Tidiness
2	Twitter archive enhanced	deleting rating_numerator which is very high (greater than 20)	Quality
3	Twitter archive enhanced	deleting rating_numerator which is less than 10	Quality
4	Twitter archive enhanced	deleting rows which do not contain 'expanded urls'	Quality
5	Twitter archive enhanced	change 'rating_numerator' to more expressive name 'rating_out_of_ten'	Quality
6	Twitter archive enhanced	tweet_id type must be a string not an integer	Quality
7	Twitter archive enhanced	Multiple dog stages : cleaning and rearranging the dog stages and merging into one column	Tidiness
8	Image prediction	tweet_id type must be a string not an integer	Quality
9	Image prediction	The columns ['jpg_url' , 'p1_dog' , 'p2_dog' , 'p3_dog'] will be removed	Quality
10	tweet_json	friends_count column should be deleted it has only one repititive irrelevent value	Tidiness
11	tweet_json	retweet status should be deleted	Tidiness
12	tweet_json	source column should be deleted	Tidiness
13	Twitter archive enhanced	columns needed to be deleted ['in_reply_status_id' , 'in_reply_to_user_id' , 'retweet_status_id' , 'retweeted_status_user_id' , 'retweeted_status_timestamp' , 'rating_denominator']	Quality
	Twitter archive enhanced	Decimal dog ratings	Quality

3.Cleaning Data

I used my knowledge of python and searching over the internet i.e. google, stackoverflow, pandas documentations... etc. for references and possible guidance to resolve the above mentioned issues to the best of my knowledge. There was lot trial and error for difficult cases where regular expressions had to be used but at the same time some things for instance dropping the not so useful columns was pretty straight forward.

Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it. Finally, once the data was ready I analyzed it using visualizations as document in act_report.pdf